DOCUMENT RESUME

ED 441 036 TM 030 835

AUTHOR Yang, Wen-Ling
TITLE Analysis of Item Ratings for Ensuring the Procedural Validity of the 1998 NAEP Achievement-Levels Setting.
SPONS AGENCY ACT, Inc., Iowa City, IA.; National Assessment Governing Board, Washington, DC.
PUB DATE 2000-04-00
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
CONTRACT ZA97001001
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; Academic Standards; Civics; *Evaluation Methods; *National Competency Tests; *Scoring; *Test Items; *Validity
IDENTIFIERS *National Assessment of Educational Progress; *Rater Effects; Standard Setting

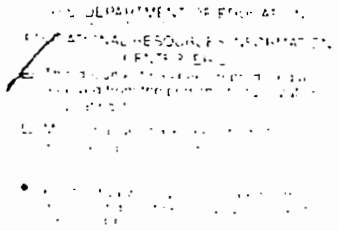
ABSTRACT

The Achievement-Levels Setting (ALS) process for the National Assessment of Educational Progress (NAEP) resulted in numerical cutscores on the NAEP score scale representing the performance standards for three achievement levels: Basic, Proficient, and Advanced. This paper focuses on an important, but less researched, aspect of the standard setting process using an item-rating approach--the patterns and changes in panelists' item ratings. Instead of analyzing the cutpoints based on item ratings, this paper presents analyses of intrarater consistency in item ratings within and across rounds of ratings, reasonableness of panelists' item ratings, and changes in ratings, in relation to cutpoints, across rounds, using the NAEP Civics examination as an example. For the 1998 achievement level setting process, the NAEP incorporated "Reckase" charts to provide extensive and easy-to-understand feedback regarding the consistency of the panelists' own ratings in the Item Response Theory context. Although the impact of Reckase charts on improving intrarater consistency could not be examined directly, Reckase generally indicated that they found the charts to be useful and informative. The analysis findings summarized in this paper are regarded as evidence of the procedural validity for the 1998 ALS process and the safeguard for the validity of the 1998 ALS outcomes. Generally, panelists were able to improve their item ratings over time. The intervention aimed at improving intrarater consistency seemed to be effective, but it was not dominating in driving panelists' subsequent item ratings. (Contains 3 tables, 4 figures, and 13 references.) (SLD)

Analysis of Item Ratings for Ensuring the Procedural Validity of The 1998 NAEP Achievement-Levels Setting

By

Wen-Ling Yang
Educational Testing Service



Wen-Ling Yang

Paper prepared for a symposium session at the annual meeting of the American Educational Research Association, New Orleans, April 2000.

The research reported in this paper was supported by contract ZA97001001 between ACT and the National Assessment Governing Board.

Analysis of Item Ratings for Ensuring the Procedural Validity of The 1998 NAEP Achievement-Levels Setting

Introduction

The Achievement-Levels Setting (ALS) process for the National Assessment of Educational Progress (NAEP) resulted in numerical cutscores on the NAEP score scale representing the performance standards for three achievement levels: Basic, Proficient, and Advanced. The purpose for establishing the achievement-levels cutpoints is to provide data on the proportion of students achieving each achievement level to improve the reporting of the NAEP assessment results. The 1998 NAEP ALS panelists received training, went through three rounds of ratings and feedback, and made recommendations for the achievement-levels cutpoints under the guidance of the Achievement Levels Descriptions (ALDs). The ALDs were developed and approved by the National Assessment Governing Board (NAGB) to provide operational definitions for the performance standards of the NAEP. Therefore, the validation issue relevant to the ALS process is whether the numerical achievement-levels cutpoints adequately represent the ALDs or the NAEP performance standards for the three achievement levels. If students performing at or above an achievement level determined by the ALS process were able to do what the ALDs described for that achievement level, the ALS outcomes would be concluded to have validity.

The NAEP assessment does not produce student scores at individual level due to its sampling design. As a result, it is impossible to validate the NAEP ALS outcomes directly. However, components of the ALS process can be carefully evaluated to ensure that the implementation of the ALS process meets the procedural requirements of the ALS design. The ALS process has been carefully researched and planned. Satisfactory implementation of the ALS process following its rigorous study design is fundamental to the procedural validity of the NAEP ALS, and hence the validity of the ALS outcomes. Much evidence has been collected for the successful implementation of the 1998 NAEP ALS. They include:

- Agreement of the cutpoints computed for the two item-rating groups.
- Similarities between the two rating groups on the set of common rating items.
- Agreement of the cutpoints computed for different item types.
- Similarities among various demographic groups of panelists in their cutpoints.
- Interrater consensus for panelists' cutpoints within rounds of ratings.
- Changes in individual cutpoints and group cutpoints across rounds.
- Evaluation results from various process and outcome evaluation questionnaires administered throughout the ALS process, indicating the extent to which panelists understood the standard-setting process and how confident they were for their ratings and achievement-levels cutpoints.

This paper focuses on an important but less researched aspect of the standard setting process using item-rating approach—the patterns and changes in panelists' item ratings. Instead of analyzing the resulting cutpoints based on item ratings, this paper presents these analyses:

- Intrarater consistency in item ratings within and across rounds of ratings.
- Reasonableness of panelists' item ratings.

- Changes in item ratings, in relation to cutpoints, across rounds.

We hope to gain insight for the validity of the ALS outcomes by investigating whether a panelist was able to yield quality ratings for items during the ALS process, and whether the study intervention was effective in improving the consistency of panelists' item ratings. In addition, changes in ratings over time were analyzed to study the improvement of panelists' item ratings from round to round. These analysis results should inform us about the adequacy of the implementation of the ALS process.

Analysis of Intrarater Consistency

After a series of field trials (Loomis, et al., 1999), pilot studies (Loomis, 1998a; Loomis, 1998b) and rigorous evaluation of the available methodologies for setting achievement standards for the NAEP, ACT decided to employ the modified-Angoff method (Angoff, 1971; Jaeger, 1989; ACT, 1993) and the Mean Score Estimation method (ACT, 1994) for the 1998 NAEP ALS process. These two chosen methods were developed for setting achievement-levels cutpoints for the dichotomously and polytomously scored items respectively. Both of them are based on the analytic item-by-item rating approach, and the cutpoint estimation procedure takes into account the Item Response Theory (IRT) context for the NAEP. Therefore, the NAEP achievement-levels cutpoints were estimated using panelists' item ratings assuming that the panelists were able to produce reasonable ratings consistent across items, regardless of item content or type. Ideally, for each item and for each achievement-level cutscore, the ALS panelists' item ratings should be consistent with the IRT calibration outcome for the NAEP. It is desired that for each item, the proportion of students at each achievement level resulted from panelists' item ratings is not very different from students' actual performance on the NAEP.

Given the item-by-item nature of the rating process, intrarater consistency is desired for the ALS panelists' item ratings. For the 1998 ALS, ACT incorporated an innovative process component—the Reckase Charts—to inform the ALS panelists of their rating consistency within rounds of ratings in a timely manner (Reckase, 1998a; Reckase, 1998b). The Reckase Charts also provided valuable rating information within and across rounds of item ratings for various research purposes. The purpose and design of the Reckase Charts are summarized below. Also presented are the possible impacts of the Reckase Charts on improving intrarater consistency and on setting the ALS cutpoints.

The Reckase Charts for Improving Intrarater Consistency

Intrarater consistency across items is an important requirement for effective rating process for setting achievement standards. Low intrarater consistency for a panelist indicates poor quality of ratings. It also implies that the panelist may have misunderstood the achievement-levels descriptions or the rating techniques. The Reckase Charts were introduced to the ALS panelists to provide extensive and easy-to-understand feedback information regarding the consistency in panelist' own ratings in the IRT context. Ideally, the charts should improve panelists' rating consistency across items within round over time, regardless of item type, content or difficulty level.

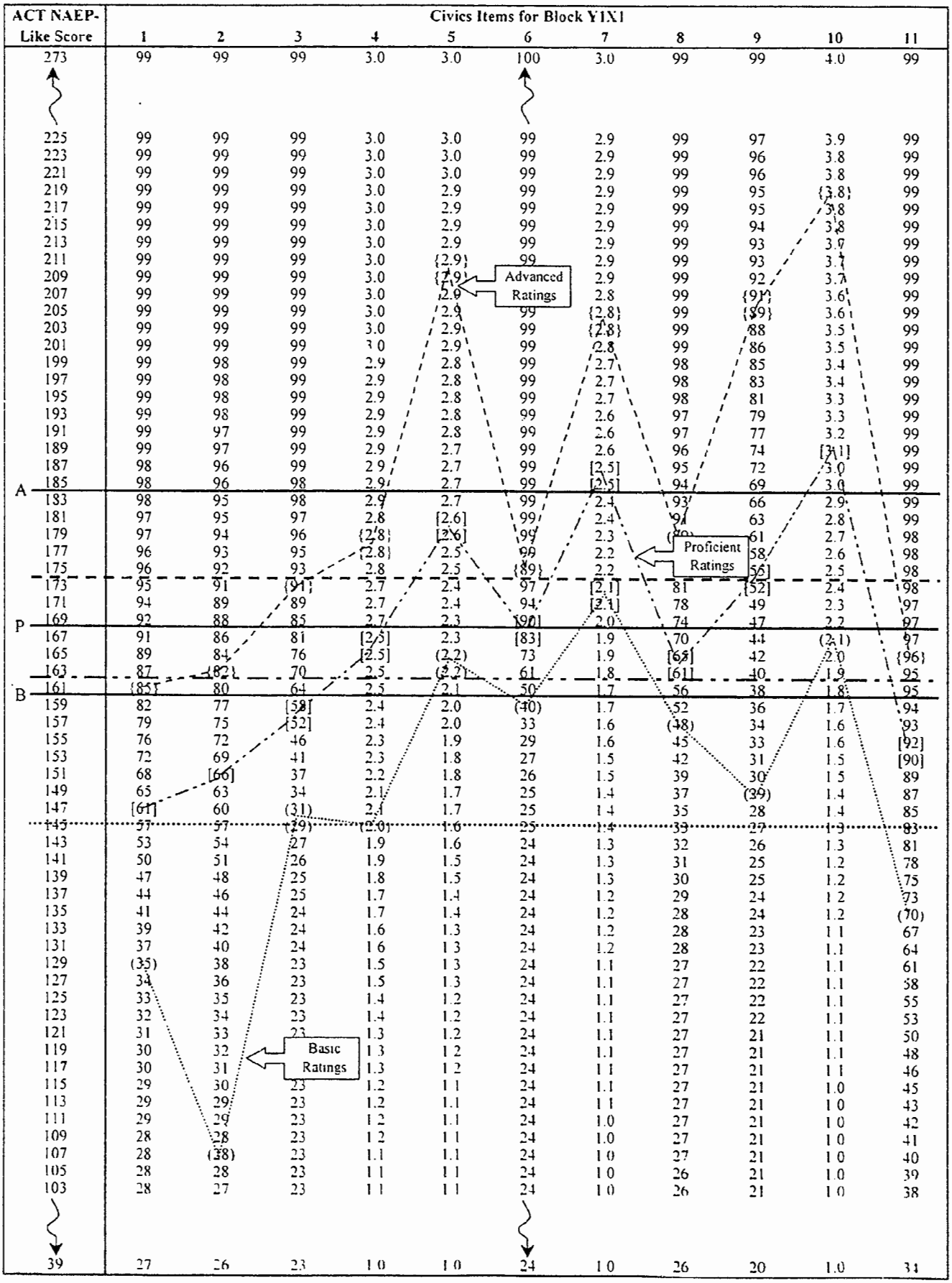
A Reckase Chart is essentially a numerical representation of the item characteristic curves for a set of rating items. An example Reckase Chart is presented in Figure 1. The numerical entries (values of probabilities or expected scores) in the body of the Reckase chart were generated by the 3-PL IRT model and the Generalized Partial Credit IRT model (Muraki, 1992) for dichotomously and polytomously scored items respectively. These numerical entries for individual rating items were arranged in columns side by side, under the heading of item numbers. In the far left column of the chart are the discrete ACT NAEP-Like Scale scores in descending order, which correspond to the probability or expected score for each of the rating. For individual panelists, item ratings from previous round are electronically marked for each of the three achievement levels, so are the estimated individual cutpoints and the grade-level cutpoints. Using the chart, a panelist can locate his/her item ratings from previous round of ratings and find the corresponding ACT NAEP-Like scale score for that item. The pattern of their ratings across items thus becomes visually clear. The panelists should be able to inspect their ratings for each item with respect to their own cutscore and the grade level cutscore, regardless of item type or content.

The impact of the Reckase Charts on improving intrarater consistency of item ratings cannot be directly examined because of the non-experimental design of the ALS study. However, the ALS panelists generally indicated in their response to the evaluation questionnaires that the Reckase Charts had being very useful and informative in helping them making adjustments to their item ratings. A substantial proportion of panelists further pointed out that they had relied more on the Reckase Charts to adjust their item ratings for the subsequent round than the other types of feedback (Hanick, 1999a). In general, a panelist's item ratings were more similar to actual student performance after the panelist reviewed the Reckase Charts than before the charts were presented to the panelist (Hanick, 1999b; Hanick, 1999c). A visual inspection on the individually customized Reckase Charts for various rounds of ratings also showed that the intrarater consistency generally improved across rounds. The changes in item ratings from round to round will be discussed later, which should also provide indications of the improvement in intrarater consistency.

Possible Intervention of Reckase Charts on Item Ratings

The Reckase Charts were first introduced to the panelists after the first round of ratings, along with other feedback material. After round 2 ratings, the panelists received their individually customized Reckase Charts again and worked with the charts in deriving their round 3 ratings. To study the influence on panelists' item ratings, possibly due to the intervention of the Reckase Charts and the other feedback, the ALS panelists' round 3 actual ratings were compared to their "expected" round 3 ratings. The "expected" ratings were derived from the panelists' round 2 cutpoints. Specifically, a panelist's round 2 cutpoint for each achievement level was expanded to obtain the "expected" ratings for the set of rating items. These "expected" item ratings could be found in a row of probability/expected scores on the Reckase Charts that corresponded to the panelist's round 2 cutpoint on the ACT NAEP-Like scale. The panelists should have become more skillful in rating items after two rounds of ratings and thus their ratings could have been more reliable over time. Therefore, we chose to focus on the last round of ratings for this analysis, instead of using ratings from earlier stage.

Figure 1. Sample Reckase Chart for Panelist *k*



ACT NAEP-Like Score	Civics Items for Block Y1X1										
	1	2	3	4	5	6	7	8	9	10	11
273	99	99	99	3.0	3.0	100	3.0	99	99	4.0	99
225	99	99	99	3.0	3.0	99	2.9	99	97	3.9	99
223	99	99	99	3.0	3.0	99	2.9	99	96	3.8	99
221	99	99	99	3.0	3.0	99	2.9	99	96	3.8	99
219	99	99	99	3.0	2.9	99	2.9	99	95	{3.8}	99
217	99	99	99	3.0	2.9	99	2.9	99	95	3.8	99
215	99	99	99	3.0	2.9	99	2.9	99	94	3.8	99
213	99	99	99	3.0	2.9	99	2.9	99	93	3.7	99
211	99	99	99	3.0	{2.9}	99	2.9	99	93	3.7	99
209	99	99	99	3.0	{2.9}	99	2.9	99	92	3.7	99
207	99	99	99	3.0	2.9	99	2.8	99	{91}	3.6	99
205	99	99	99	3.0	2.9	99	{2.8}	99	{89}	3.6	99
203	99	99	99	3.0	2.9	99	{2.8}	99	88	3.5	99
201	99	99	99	3.0	2.9	99	2.8	99	86	3.5	99
199	99	98	99	2.9	2.8	99	2.7	98	85	3.4	99
197	99	98	99	2.9	2.8	99	2.7	98	83	3.4	99
195	99	98	99	2.9	2.8	99	2.7	98	81	3.3	99
193	99	98	99	2.9	2.8	99	2.6	97	79	3.3	99
191	99	97	99	2.9	2.8	99	2.6	97	77	3.2	99
189	99	97	99	2.9	2.7	99	2.6	96	74	[3.1]	99
187	98	96	99	2.9	2.7	99	[2.5]	95	72	3.0	99
185	98	96	98	2.9	2.7	99	[2.5]	94	69	3.0	99
183	98	95	98	2.9	2.7	99	2.4	93	66	2.9	99
181	97	95	97	2.8	[2.6]	99	2.4	91	63	2.8	99
179	97	94	96	{2.8}	[2.6]	99	2.3	{89}	61	2.7	98
177	96	93	95	{2.8}	2.5	99	2.2	86	58	2.6	98
175	96	92	93	2.8	2.5	{89}	2.2	83	56	2.5	98
173	95	91	{91}	2.7	2.4	97	[2.1]	81	[52]	2.4	98
171	94	89	89	2.7	2.4	94	[2.1]	78	49	2.3	97
169	92	88	85	2.7	2.3	[90]	2.0	74	47	2.2	97
167	91	86	81	[2.6]	2.3	[83]	1.9	70	44	(2.1)	97
165	89	84	76	[2.5]	(2.2)	73	1.9	[65]	42	2.0	{96}
163	87	82	70	2.5	(2.2)	61	1.8	[61]	40	1.9	95
161	{85}	80	64	2.5	2.1	50	1.7	56	38	1.8	95
159	82	77	[58]	2.4	2.0	(40)	1.7	52	36	1.7	94
157	79	75	[52]	2.4	2.0	33	1.6	(38)	34	1.6	93
155	76	72	46	2.3	1.9	29	1.6	45	33	1.6	[92]
153	72	69	41	2.3	1.8	27	1.5	42	31	1.5	[90]
151	68	[66]	37	2.2	1.8	26	1.5	39	30	1.5	89
149	65	63	34	2.1	1.7	25	1.4	37	(39)	1.4	87
147	[61]	60	(31)	2.1	1.7	25	1.4	35	28	1.4	85
145	57	57	(29)	(2.0)	1.6	25	1.4	33	27	1.3	83
143	53	54	27	1.9	1.6	24	1.3	32	26	1.3	81
141	50	51	26	1.9	1.5	24	1.3	31	25	1.2	78
139	47	48	25	1.8	1.5	24	1.3	30	25	1.2	75
137	44	46	25	1.7	1.4	24	1.2	29	24	1.2	73
135	41	44	24	1.7	1.4	24	1.2	28	24	1.2	(70)
133	39	42	24	1.6	1.3	24	1.2	28	23	1.1	67
131	37	40	24	1.6	1.3	24	1.2	28	23	1.1	64
129	(35)	38	23	1.5	1.3	24	1.1	27	22	1.1	61
127	34	36	23	1.5	1.3	24	1.1	27	22	1.1	58
125	33	35	23	1.4	1.2	24	1.1	27	22	1.1	55
123	32	34	23	1.4	1.2	24	1.1	27	22	1.1	53
121	31	33	23	1.3	1.2	24	1.1	27	21	1.1	50
119	30	32	23	1.3	1.2	24	1.1	27	21	1.1	48
117	30	31	23	1.3	1.2	24	1.1	27	21	1.1	46
115	29	30	23	1.2	1.1	24	1.1	27	21	1.0	45
113	29	29	23	1.2	1.1	24	1.1	27	21	1.0	43
111	29	29	23	1.2	1.1	24	1.0	27	21	1.0	42
109	28	28	23	1.2	1.1	24	1.0	27	21	1.0	41
107	28	(28)	23	1.1	1.1	24	1.0	27	21	1.0	40
105	28	28	23	1.1	1.1	24	1.0	26	21	1.0	39
103	28	27	23	1.1	1.1	24	1.0	26	21	1.0	38
39	27	26	23	1.0	1.0	24	1.0	26	20	1.0	34

Analysis results have indicated that panelists' cutpoints usually did not vary much from round 2 to round 3. Thus, it may be reasonable to assume that panelists were quite satisfied with their round 2 cutpoints and only wanted to make minor adjustments to their round 3 item ratings to reflect the slight change in the cutpoints. Under such circumstances, if the Reckase Charts were informative to the panelists, panelists were likely to yield ratings for round 3 that were only slightly different from their expected ratings associated with their round 2 cutpoints. It is because when panelists reviewed their round 2 rating outcomes with the Reckase Charts, the rating values associated with their round 2 cutpoints could easily catch their attention and influence their round 3 item ratings. Therefore, a large discrepancy between the round 3 actual ratings and "expected" ratings can imply that the Reckase Charts did not have substantial didactic impact on panelist's subsequent item ratings. To the opposite, a small discrepancy between the round 3 actual ratings and "expected" ratings could indicate that the Reckase Charts had didactic impact on panelist's subsequent item ratings, assuming that the impacts from the other feedback were minimal.

For the 1998 Civics NAEP

The direction and magnitude of discrepancies between panelists' round 3 actual ratings and expected ratings were analyzed by grade and by achievement level for each type of items. Also examined were analysis outcomes by rating group and by individual panelists. Summary tables and graphs were constructed to summarize percentages of items for which actual ratings were higher than, lower than, or equal to expected ratings for each grade at each achievement level. For the 1998 Civics NAEP, the overall patterns of rating discrepancies were different for various item types. The implication is that the intervention effect of the Reckase Charts, if present, might vary with the types of rating items.

In general, there was only a small proportion of multiple-choice items for which round 3 actual ratings coincided with expected ratings. For all achievement levels and for all grades, most panelists had more ratings below the expected line on the Reckase Charts from round 2. It suggested that panelists generally lowered their item ratings from round 2 to round 3 for the multiple-choice items. The exception is for grade 4 at the basic level, where there were more items for which ratings were above the expected line. Also for the short-response items, only a small proportion of items for which round 3 actual ratings and expected ratings overlapped. For grade 4, most panelists had ratings above the expected line across achievement levels. For both grade 8 and grade 12 at the advanced levels, panelists often had ratings below the expected line. For the extended-response items, round 3 actual ratings generally agreed more with the expected ratings than the other two types of items. Across achievement levels and grades, most item ratings were above the expected line, though.

Ratings for different types of items were transformed to a proportional magnitude for comparison purposes. For short-response items, there were larger discrepancies than the other two types of items across grades and achievement levels. In addition, for the basic and the advanced levels, the average discrepancies on extended-response items were larger than multiple-choice items across grades.

For the 1998 Writing NAEP

Except for two panelists, the 1998 writing ALS panelists had a large proportion of items for which their ratings were different from the expected ratings. It suggested that most of the panelists made adjustments for their round 3 item ratings such that their round 3 cutpoints deviated from their round 2 cutpoints. On average, grade 4 had the highest percentages of items for which actual ratings were different from expected ratings among the three grades. It indicated that the grade 4 panelists generally made more adjustments for their round 3 item ratings than the other two grades, such that their round 3 item ratings were more different from round 2 cutpoints than the other grades. The magnitude of the discrepancy, however, was not large for all three grades. That is, round 3 actual ratings for all three grades were not very different from expected ratings in magnitude.

Cautions for the Use of the Reckase Charts

Various hypotheses were formulated to examine the impact of the Reckase Charts in informing the ALS panelists about the intrarater consistency in their item ratings within rounds and across rounds. The customized Reckase Charts reviewed by individual panelists during the rating process, data from the evaluation questionnaires, and panelists' rating data and cutpoint estimation outcomes were examined to test these hypotheses in a fashion of triangulation. This paper presented the analysis results for the hypotheses of the most interest. Although the Reckase Charts seemed to be informative to the ALS panelists during the item rating process, analysis results generally did not reveal a clear pattern for the influence of the Reckase Charts.

The type and timing of feedback are critical to the standard setting process. The feedback is used to help panelists understand the capabilities of students and to gain an understanding of the relationship between their item ratings and the cutpoints on the IRT-based reporting scale. Any feedback should not become a single deciding factor for panelists' consideration and decision about the cutpoints. Therefore, despite that the Reckase Charts provide consistency information for the panelists to help them adjust their item ratings to be more in line with the item characteristic curves (ICC), panelists' item ratings should not be driven by the ICC embedded in the Reckase Charts. The analysis findings summarized in this section generally supported the notion that the impact of the Reckase Charts on panelists item ratings might not be strong or obvious. Therefore, the concern about the dominating impact from the Reckase Charts on the last round of item ratings was not serious.

Extreme Values for the Last Round of Item Ratings

Before beginning their ratings for items, the 1998 ALS panelists received three days of training and a variety of information regarding setting achievement levels for the NAEP. After each round of ratings, they also received feedback for their item ratings including individual cutpoint and grade-level cutpoints, the rater location data, the consequences data, and the individually customized Reckase Charts. Given the practice in rating items over time and the informative feedback, it is reasonable to expect that panelists would produce less unreasonable ratings for items across rounds. Particularly for the last round of ratings, they should not be

producing many extreme values for item ratings. It is of research interest to identify panelists who still made extreme item ratings at the final stage of the ALS process. Special attention should be paid to these panelists' rating patterns to understand the cause of their extreme ratings.

Criterion for Identifying Extreme Item Ratings

The Reckase charts provide item characteristics information based on the IRT scaling of the NAEP to the panelists. The scores at the two ends of the ACT NAEP-Like scale on the Reckase Charts represent extreme scores for NAEP items for which students rarely get. Therefore, these extreme scores were used as the criterion for identifying extreme item ratings for the 1998 Civics and Writing ALS panelists. Specifically, to determine whether panelists' last-round item ratings were extreme, item ratings were first transformed to ACT NAEP-Like Scale scores. Then, the transformed scores were compared to the two criterion scores—the high-end and the low-end scores on the ACT NAEP-Like scale presented on the Reckase Charts. If the value for a transformed item rating was larger than the high-end ACT NAEP-Like score for that item, the rating was considered extremely large. If the value was smaller than the low-end ACT NAEP-Like score, the rating was regarded as unreasonably small.

Extreme Ratings for the Last Round of the 1998 Civics NAEP

The Civics ALS panelists who had extreme ratings were identified using the above criterion for all three grades. To further investigate the degree of these extreme ratings, the raw rating values for the identified extreme ratings were compared to the upper or lower bound of the ACT NAEP-Like score for each item, as shown in the body of the Reckase Charts. The discrepancy between the raw rating value and the upper or lower bound of the rating scale for each item was informative in reasoning why panelists produced extreme rating values at the last round of the ALS process. On average, the discrepancy between the raw values for the extreme ratings and the upper/lower bounds of the rating scales for items was not large. It indicated that these extreme rating values were not too far apart from the reasonable range for item ratings.

It is found that 86% of the total extreme ratings was set for the Basic achievement-level cutpoint. A close inspection on the items received extreme ratings for the Basic level cutpoint further revealed that most of these items had a rather great chance level for being answered correctly. The great chance levels helped explain why some of the ALS panelists yielded rating values that were considered extreme for the Basic level cutpoint. For items with great chance level, it is more likely for panelists to give rating values for the Basic level cutscore that fall below the chance level.

Information about the type (teacher, non-teacher educator, or general public) for the panelists identified to have extreme ratings was obtained for analysis purposes. Results of the analyses by panelist type and the analyses at the individual level are summarized below. In addition, analyses of extreme values at grade level were summarized in a table. Table 1 shows the number of panelists with extreme item ratings. It also displays the number of items for which panelists assigned extreme ratings.

Analysis by Panelist Type

Roughly, 55% of all the Civics panelists were teachers, 15% of them were non-teacher educators, and the other 30% represented general public. Important findings for analyses of extreme ratings by panelist type by grade include:

- For grade 12, none of the six general-public panelists (0%) had extreme round 3 item ratings, and only one of the four non-teacher educators (25%) had extreme ratings. This non-teacher educator had ratings for the Basic level cutpoint for five items that were slightly lower than the expected minimum ratings. Nine of the 17 teachers (53%) had at least one extreme item ratings. In short, for grade 12, more teachers produced extreme ratings than the other two types of panelists.
- For grade 8, none of the four (0%) non-teacher educators had extreme ratings. The proportion of teachers who made extreme ratings was smaller than the proportion of the general-public panelists. It is found that five of the nine (56%) general-public panelists and seven of the 16 (44%) teachers had extreme ratings. However, all the extreme ratings were set for the Basic level cutpoint and most of them were slightly lower than the expected minimum ratings for various items.
- For grade 4, only one of the four (25%) non-teacher educators had extreme ratings, but four of the eight (50%) general-public panelists and six of the 19 (32%) teachers did. The non-teacher educator only had extreme values for two items and they were not very serious. Overall, as in grade 8, the proportion of teachers with extreme ratings was smaller than the proportion of general-public panelists. Also, for both grade 8 and grade 4, the non-teacher educators often had less extreme item ratings.

Special Individual Cases

Some interesting analysis outcomes at individual level are highlighted below:

- One grade-4 general-public panelist had ratings that were lower than the expected minimum rating for both the Basic and Proficient level cutpoints for the same item. A possible reason for the extreme ratings is that the panelist considered the item fairly difficult such that even proficient-level students would not have a great chance for answering it correctly.
- One grade-4 teacher panelist had the same extremely high rating (100%) to the three achievement level cutpoints for the same multiple-choice item. A possible explanation is that this panelist regarded the dichotomously scored item extremely easy and non-discriminating.
- For grade 8, all of the extreme ratings were produced for the Basic achievement-level cutpoint. It suggested that some grade 8 panelists (12/29) tended to set low standards for the Basic achievement level for some items. As explained earlier, most of these items were found to have a rather great chance level for being answered correctly. It is why these grade-8 panelists were more likely to yield ratings that fall below the chance level for the Basic level cutscore.
- There was one grade 12 panelist who had extremely low ratings for both the basic and the Proficient levels for the same item. Another grade 12 panelist not only had an extremely low rating for a multiple-choice item for the Proficient-level cutpoint, the rating was even lower than the rating for the Basic-level cutpoint for the same item. Note that the data entry program used for the ALS process typically checked for such illogical ratings and panelists were given the opportunity to correct errors in their ratings. However, this panelist's rating

error was not corrected for the Civics ALS because the panelist was not available for make any correction for that item and the limited time for on-site computation prohibited the long wait for that panelist.

Grade-Level Analysis

For each round, there were 21,183 rating values produced for all the items rated by the 1998 Civics ALS panelists from all three grades for all three achievement levels. Table 1 shows the total number of items rated by the ALS panelists by grade and group. It also indicates the number of items rated by item type.

Using the criterion for identifying extreme ratings described above, Table 2 shows that there were a total of 100 counts of extreme ratings for the last round of ratings. The number of extreme ratings was less than 0.5% of the total rating values produced across grades for the 1998 Civics ALS. Some panelists had extreme ratings for more than one item, and some items received extreme ratings from more than one panelist. Among the three grades, grade 8 panelists had the most number of extreme ratings (44/100) and grade 4 panelists had the least number of extreme ratings (23/100). Across grades, there were about the same numbers of panelists who had at least one extreme item rating (11 panelists for grade 4, 12 for grade 8, and 10 for grade 12). However, for grade 8, there were five panelists who had at least five extreme item ratings. For grade 4, only one panelist produced at least five extreme ratings. For grade 12, there were three similar panelists.

Across grades, there were a total of 58 items that received extreme ratings from at least one panelist, as shown in Table 2. For grade 8, there were 24 items that had extreme ratings. For grade 4, there were only 13 such items; and for grade 12, there were 21 items. In addition, 11 items among these 58 items across grades had extreme ratings from at least three panelists. There were five such items for grade 8, four for grade 12, and two for grade 4.

Extreme Ratings for the Last Round of the 1998 Writing NAEP

Overall, the last round of ratings produced 3,168 rating values for all the items (all constructed-response) rated by the 1998 Writing ALS panelists from all three grades for all three achievement levels. Table 3 shows the total number of the Writing ALS panelists, and the number of items rated by these panelists by rating group by grade.

Using a similar criterion for identifying extreme ratings, no extreme round-3 item ratings were found for the 1998 Writing ALS panelists. For all three grades, all of the transformed item ratings were within the ACT NAEP-Like score range specified on the Reckase Charts.

Analyses of Ratings Changes from Round to Round

The average percentage of items for which ratings were changed (raised or lowered) or unchanged from round to round were studied. Important findings are summarized in this section for the 1998 Civics NAEP and the 1998 Writing NAEP respectively. An innovative data plot

Table 1. Number of Items Rated by the 1998 Civics ALS Panelists

Grade	Group	# of Panelist	# of Items			
			Total	Multiple-Choice	Short Constructed-Response	Extended Constructed-Response
4	A	16	58	44	10	4
	B	15	59	46	10	3
8	A	15	93	77	12	4
	B	14	93	76	12	5
12	A	14	94	77	14	3
	B	13	95	77	14	4

Table 2. Summary of Extreme Item Ratings for the Last Round of Item Ratings of the 1998 Civics ALS Process

Grade	Total Count of Items with Extreme Ratings	# of Panelists with at least **ONE** Extreme Item Rating	# of Panelists with at least **FIVE** Extreme Item Ratings	# of Items with Extreme Ratings from at least **ONE** Panelist	# of Items with Extreme Ratings from at least **Three** Panelists
4	23	11	1	13	2
8	44	12	5	24	5
12	33	10	3	21	4
Total	100	33	9	58	11

Table 3. Number of Items Rated by the 1998 Writing ALS Panelists

Grade	Group	# of Panelists	# of Items (All Constructed-Response)
4	A	14	12
	B	15	12
8	A	15	12
	B	15	12
12	A	14	12
	B	15	12

developed for summarizing rating changes for individual panelists, in relation to their individual cutpoints and the grade-level cutpoints, is also introduced in this section.

Summary for the 1998 Civics NAEP

It is found that, across grades and achievement levels, panelists consistently made changes on fewer items from round 2 to round 3 than from round 1 to round 2. Both rating-group data and individual panelist data were examined to study patterns of rating changes. The individual data showed that:

- Most of the Civics panelists raised their ratings on more items than lowering their item ratings from round 1 to round 2.
- Four panelists only raised their item ratings without lowering ratings for any items.
- From round 2 to round 3, four panelists did not change any of their item ratings, and many others only changed their ratings for a very small proportion of items.
- From round 2 to round 3, there were still five panelists who made rating changes to more than half of the items. Mainly, they raised ratings for these items for all the achievement levels. The data for rating changes in magnitude further showed that the rating changes made by these five panelists from round 2 to round 3 were generally larger than the changes of the other panelists. However, the magnitude of their changes in ratings was often quite small—less than 10%.

The average absolute amount of rating changes from round to round was computed by grade for each achievement level. The magnitude data was analyzed by item type (multiple-choice, short-response, and extended-response). The magnitude of rating changes for the polytomous items was transformed to a proportional magnitude to facilitate comparisons across item type. As expected, panelists in general had smaller rating changes in magnitude from round 2 to round 3 than from round 1 to round 2 for all item types. Individual panelist rating data showed that:

- From round 1 to round 2, three panelists did not have any rating changes for short-response and extended response items, and their changes for multiple-choice items were relatively small.
- Eight panelists had relatively large changes (within rating group) from round 1 to round 2 Among them, two panelists not only changed their ratings extensively but also sharply from round 1 to round 2.
- While 47 panelists did not change their ratings on extended-response items across achievement levels and 28 did not change on short-response items, only five panelists carried over their ratings from round 2 to round 3 for multiple-choice items. The differences across item types were similar for all three grades.
- From round 2 to round 3, the number of panelists that had relatively large changes in magnitude across achievement levels are: 5 panelists for grade 4; 3 panelists for grade 8; and 3 panelists for grade 12.

Summary for the 1998 Writing NAEP

For all grades and all achievement levels, the Writing panelists made changes on fewer items from round 2 to round 3 than from round 1 to round 2. For grade 4 and grade 12, panelists generally raised ratings on more items than lowering ratings on items. The patterns of rating changes for grade 8, however, were quite different. Except for the basic level from round 1 to round 2, the grade 8 panelists generally lowered ratings on more items than raising ratings on items. Figure 2 summaries the average percentage of items for which ratings were raised, lowered, or remained unchanged from one round. The average absolute amounts of changes from round to round in item ratings are reported by grade for each achievement level in Figure 3. The magnitude of changes for writing items, all polytomously scored, was transformed to a proportional magnitude. Figure 3 clearly shows that overall panelists made much smaller changes in magnitude for their item ratings from round 2 to round 3 than from round 1 to round 2. This pattern was consistent across all achievement levels and all grades.

Across grades from round 1 to round 2, six panelists only raised item ratings without lowering ratings for any items for all achievement levels. Four other panelists only lowered item ratings without raising any ratings. From round 2 to round 3, three panelists did not change any of their item ratings. Many others only changed their ratings for a very small proportion of items. Overall, a total of twenty panelists did not raise ratings on any items. Also, panelists made much fewer changes on their ratings from round 2 to round 3 than from round 1 to round 2.

Analyses based on RMSD

Additional to the above analyses, the magnitude of changes in item ratings for the 1998 Writing ALS was further examined using the ACT NAEP-Like scores transformed from the panelist's actual ratings. The Root-Mean-Squared Deviation (RMSD) statistic (Schmitt, Cook, Dorans, & Eignor, 1990) was computed to estimate the magnitude of changes from round to round. The RMSD statistic was computed as follows:

$$RMSD = \{[\sum n_y (\hat{x}_y - x_y)^2] / \sum n_y\}^{\frac{1}{2}},$$

where $\hat{x}_y$ and x_y are the ACT NAEP-Like scores transformed from ratings for two different rounds respectively for panelist y. Since the data analyzed were not categorical, n_y is a constant that equals one. The summation is over the number of items rated by the panelist. Analysis outcomes based on the RMSD were summarized below.

Averages of rating changes from round to round in magnitude were computed by grade for each achievement level. As found in previous analyses using item rating data on the proportional scale, panelists generally had much smaller changes in their ratings from round 2 to round 3 than from round 1 to round 2. For grade 12 at the advanced level from round 1 to round 2, the average magnitude of change seemed strikingly large. A close examination of panelists' data revealed that there were more panelists in grade 12 at the advanced level with large magnitude of rating change than the other two grades.

Figure 2. Proportions of Writing Items for which Item Ratings Changed

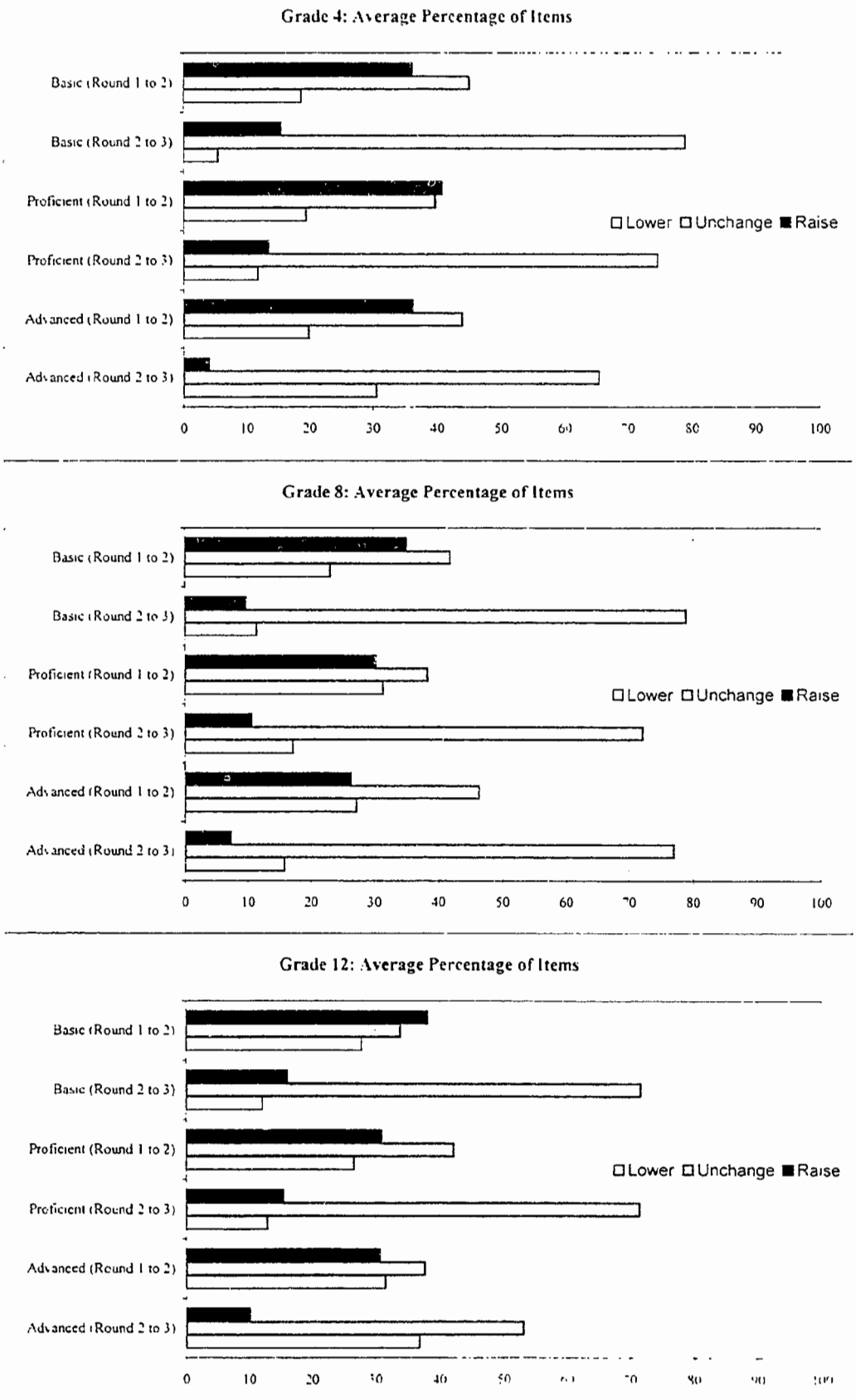
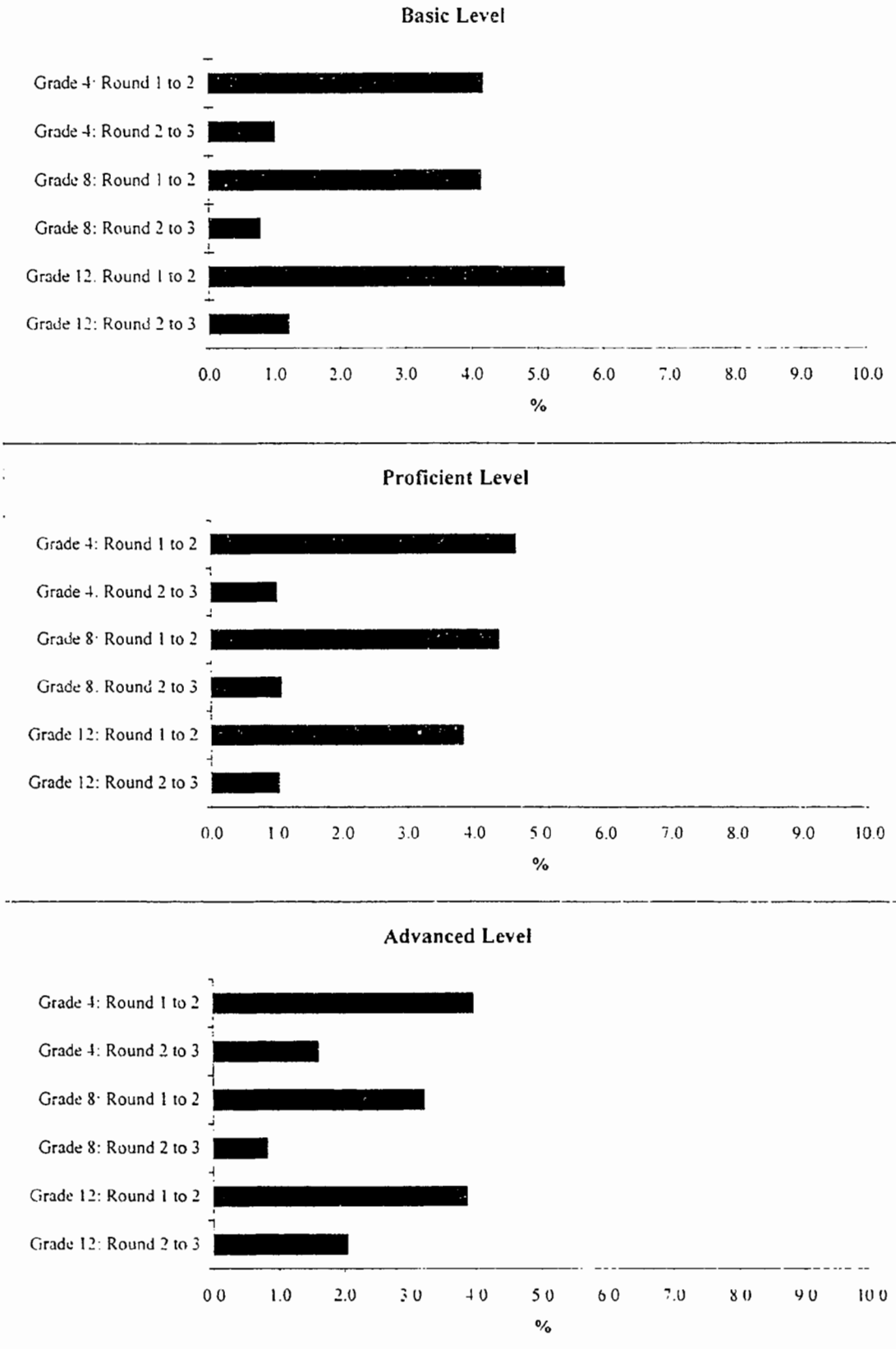


Figure 3. The Magnitude of Average Rating Changes for the 1998 Writing ALS



Analysis using RMSD was also conducted by rating groups. Note that for grade 12 at the advanced level from round 1 to round 2, one rating group had relatively large magnitude of change, compared to the other group. Panelist data indicated that for this group of panelists, there were more changes with large magnitude than the other groups. For grade 8 across achievement levels, one group had larger changes than the other group from round to round. This finding was very similar to a previous finding based on the absolute values of differences on the proportional scale. No other clear patterns were found across achievement levels for the other grades.

The results of analyses on individual data were not very different from previous analysis outcomes. From round 2 to round 3, most of the panelists made smaller rating changes than from round 1 to round 2. However, across the three achievement levels and the three grades, five panelists still made relatively large changes in their item ratings.

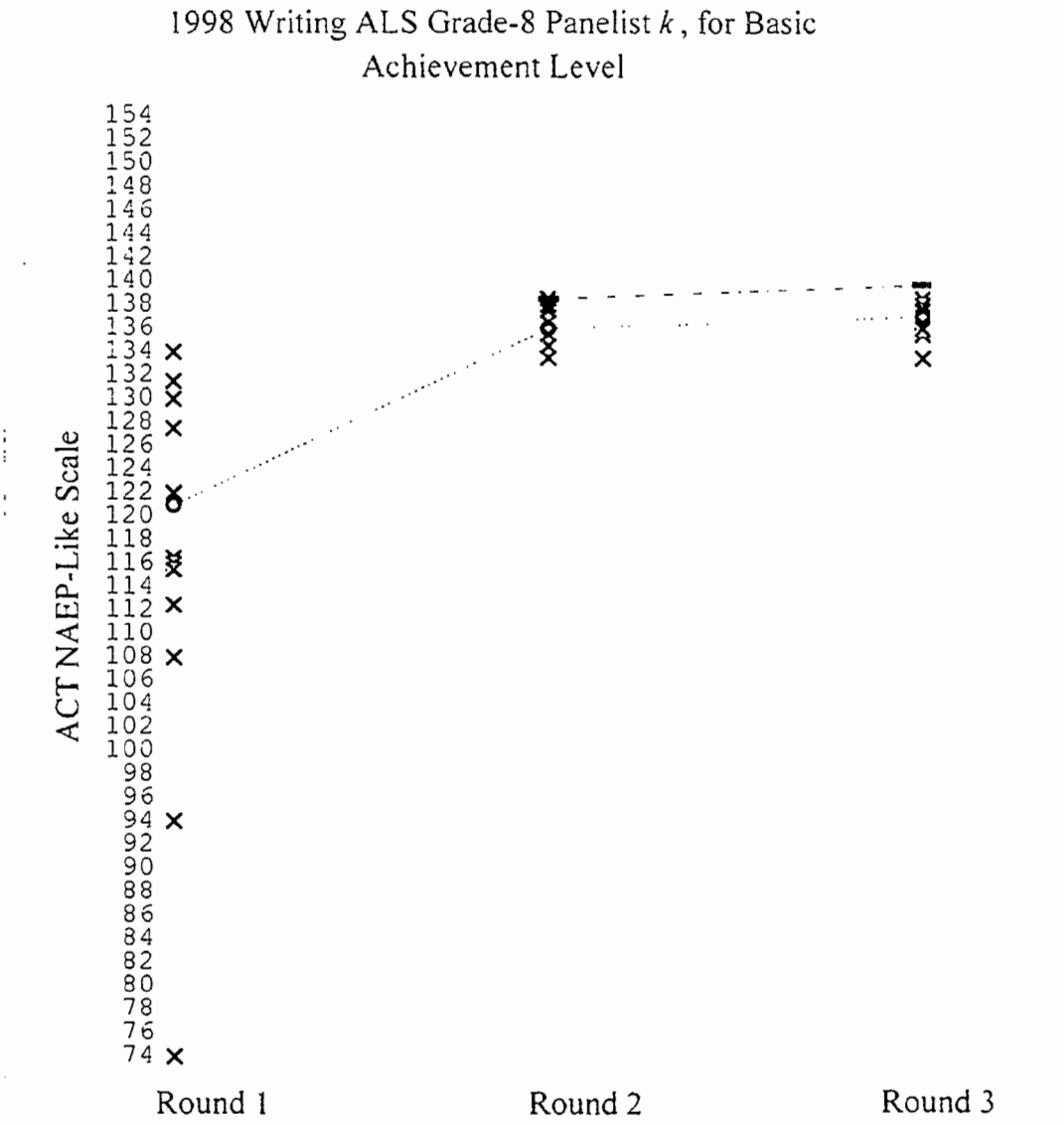
Visual Representation of Rating Changes for Individual Panelists

Various schemes were developed for summarizing changes in item ratings for individual panelists from round to round. They include several graphic representations of the rating changes for individual panelists. Among them, the one that takes into account the relations between panelist's item ratings and individual cutpoint, and the grade-level cutpoint from round to round is regarded as the most effective and most efficient. This type of graph has the advantages of consolidating a panelist's item ratings and cutpoints on the same scale, and presenting rating data from all rounds in one chart, such that the patterns of rating changes (location and spread) are visually clear and the comparisons across rounds are easy. An example plot is presented in Figure 4 to illustrate the application of the innovative graphic representation for summarizing rating changes for individual panelists from round to round.

The summary plot described above requires that panelist's item ratings be transformed to the ACT NAEP-Like Scale score, so that the rating outcomes can be compared to the resulting cutpoints on the same scale. The estimated individual cutpoint for a panelist is plotted on the same (vertical) line of the transformed item ratings for each round. The spread of item ratings around the individual cutpoint is made easy for inspection for each round. In addition, the grade-level cutpoints from previous rounds are plotted on the line of item ratings for current round. It is to help inspect the impact of the grade-level cutpoint on the next-round item ratings.

The summary plot is particularly useful for exploring item-rating data of interesting panelist cases or investigating special panelist cases. For instance, for the last two rounds of the 1998 Writing ALS process, none of the panelists was found to have item ratings with large standard deviation (greater than one standard deviation of the ACT NAEP-Like scale). For the first round of the process, however, several panelists from various grades produced item ratings for various achievement levels with large standard deviations. To examine whether the item ratings of these panelists were extreme and whether the spread of item ratings decreased over rounds, a summary plot can be constructed for each of these panelists. The plots will also provide insights for whether the estimated individual and grade-level cutpoints were influential to the panelists on their next-round item ratings.

Figure 4. Example Plot for Summarizing Panelist's Item Ratings and Cutpoint, and the Grade-Level Cutpoint across Rounds



x denotes rating on the ACT NAEP-Like scale.

o denotes individual panelist's cutpoint.

\- denotes grade-level cutpoint.

Note. For round 2, the grade-level cutpoint from round 1 is plotted; and for round 3, the grade-level cutpoint from round 2 is plotted.

Conclusions

In this study, intrarater consistency in item ratings was examined, extreme item ratings were defined and scrutinized, and the changes in item ratings from round to round were studied in relation to the resulting cutpoints. The various analysis findings summarized in this paper are regarded as the evidence of the procedural validity for the 1998 ALS process and the safeguard for the validity of the 1998 ALS outcomes. Generally, given practice and informative feedback, panelists were able to improve the quality of their item ratings over time. The study intervention aimed at improving intrarater consistency seemed to be effective but it was not dominating in driving panelists' subsequent item ratings. In addition, the analysis for item rating changes showed less and smaller changes from round 2 to round 3 than from round 1 to round 2. It indicated the increasing stability of panelists' item ratings over time.

References

ACT (1993). Setting Achievement Levels on the 1992 National Assessment of Educational Progress in Mathematics, Reading, and Writing: A technical report on reliability and validity. Iowa City, IA: Author.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.) Educational Measurement (2nd Edition). Washington, DC: American Council on Education.

Hanick, P. L. (1999a). 1998 Civics NAEP Achievement Levels-Setting meeting: Summary report of process evaluation questionnaires. Report to the Technical Advisory Committee on Standard Setting, February 18-19, 1999, Atlanta.

Hanick, P. L. (1998b). Civics pilot study process evaluation questionnaire summary report. Report to the Technical Advisory Committee on Standard Setting, September 1998, Minneapolis.

Hanick, P. L. (1998c). Writing pilot study process evaluation questionnaire summary report. Report to the Technical Advisory Committee on Standard Setting, October 1998, Detroit.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.) Educational Measurement (3rd Edition). New York: American Council on Education & Macmillan.

Loomis, S. C. (1998a). NAEP 1998 Civics ALS pilot study overview. Report to the Technical Advisory Committee on Standard Setting, September 1998, Minneapolis.

Loomis, S. C. (1998b). NAEP 1998 Writing ALS pilot study summary. Report to the Technical Advisory Committee on Standard Setting, October 1998, Detroit.

Loomis, S. C., Bay, L., Yang, W. L., & Hanick, P. L. (1999). Field trials to determine which rating method(s) to use in the 1998 NAEP Achievement Levels-Setting process for Civics and Writing. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, 1999.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16(2), 159-176.

Reckase, M. D. (1998a). Converting boundaries between National Assessment Governing Board performance categories to points on the National Assessment of Educational Progress score scale: The 1996 science NAEP process. Applied Measurement in Education, 11 (1), 9-21.

Reckase, M.D. (1998b). Setting standards to be consistent with an IRT item calibration. Iowa City, IA: ACT.

Schmitt, A. P., Cook, L. L., Dorans, N. J., & Eignor, D. R. (1990). Sensitivity of equating results to different sampling strategies. Applied Measurement in Education. 3, 53-71.